

**IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
APPLICATION FOR UNITED STATES LETTERS PATENT**

INVENTORS:

Ophir FRIEDER
David Adam GROSSMAN

TITLE:

INTRANET MEDIATOR

ATTORNEYS:

Roland W. Norris
Pauley Petersen Kinne & Fejer
2800 West Higgins Road
Suite 365
Hoffman Estates, Illinois 60195
(847) 490-1400

EXPRESS MAIL NO.: EL688022866US

MAILED: 18 April 2001

BACKGROUND OF THE INVENTION

Field of the Invention

The present invention relates to a system whereby a direct answer may be given to a specific natural language query through a convenient search of
5 structured and unstructured data sources.

Discussion of the Related Art

There are two types of digital data gathering commonly in use. One,
information retrieval, is concerned with the retrieval of information from unstructured
data sources, such as text documents, where each element of the data is not
individually defined. The user will enter "search terms" as a data query and the
unstructured data will be searched for occurrence of these terms. Results of such a
search may return the text, i.e., the data, or may, e.g., in a World Wide Web search,
only return the location, or site, of the data. The user would then need to read the text
or go to each site and locate the occurrence of the search term, which may, or may
not, be relevant to an actual question which the user wants answered. This time
15 consuming practice is commonly known as "surfing".

Information retrieval is thus not geared to efficiently provide a specific
answer to a specific question. Attempts to alleviate this problem were the subject of
U.S. Patent 6,167,370 to Tsourikov et al., which suggests giving a summary of text
20 findings as a response to a user query. But, for example, when a user wants to know
"What are the three best Sushi restaurants in Chicago?" the user does not necessarily
care to browse through text summaries, or restaurant guide web sites, which are the

likely search results of a known information retrieval search. The surfing in this context may be particularly tedious if the query is submitted to the data sources as an equally weighted string of tokens. For example, where "Chicago" is equally weighted with "Sushi" when figured into the search results, a user may wade through scores of restaurant web sites having nothing to do with Sushi eateries. Avoidance of this problem may require the user to know Boolean logic or other specific search strategy formats, and individually structure each search. The user would most often prefer just a list of three Sushi restaurants in Chicago in response to this natural language question.

The second type of digital data gathering commonly in use is the structured data source search, where highly structured data within one specific data source, usually privately owned and accessed, are searched to return a specific answer. In the past, the data sources were required to be searched one data source at a time. Integration of their individual data sources is generally performed by private business to enable answers to queries whose answers require more than one factual component. This integration is expensive and can remain underutilized for reasons such as an arcane nature of query formulation or because extensive data source knowledge may be required of the user to make a rational search selection. That is, the user may need to know where to look and how to look to expect a relevant answer.

Concurrent searching of unintegrated structured data sources, and merging of their results, to solve some of these problems, was the subject of U.S. Patent 5,995,961 to Levy, et al.

Further, additional information, beyond the specific factual components of a query, cannot be provided from the results of a data source search. For example, assume that the data source user, or searcher, wishes to know the building on the I.I.T. campus with the largest number of rooms. The user cannot expect a picture of the 5 building, or a link to a picture of the building, returned with the search results, even though the user might wish to see such a picture.

U.S. Patent No. 6,078,924 to Ainsbury et al. illustrates a technique of digital data gathering. According to this patent, the user is allowed to aggregate data found in the user's previous searches on a specific topic into a central file. This 10 central file can then be controlled from a commercial desktop computer application to facilitate searching of the data.

What is needed in the art is a system whereby the user can take advantage of both information retrieval and structured data types of digital data gathering concurrently to provide a direct answer to a specific question, and 15 preferably provide further context for that answer. It is also desirable that the query be accepted in a natural language format whereby the user needs no special skills in query formulation. It is further desirable that the query be intelligently parsed so as to weight the relevant parts of the query and that synonyms of the natural language query be provided to give a more thorough search and accurate answer. It is further 20 desirable that the answers, and any related information, be limited in number to only that required or most relevant to the query.

DEFINITIONS

“Query” refers herein to any form of searchable subject matter, and may include query tokens, or elements of a total query, whether aggregate or separate, unless otherwise limited or defined by the context of the disclosure.

5 “Data” refers herein to any form of digitally stored information, unless otherwise limited or defined by the context of the disclosure.

“Concurrent” means within the time frame between question output and answer display and does not necessarily imply that searches happen simultaneously.

10 “Direct answer” and “most likely answer” are used interchangeably herein and refer to the best available answer, whether factually based, referencing additional data, or refusal to answer, based upon the results of the data retrieved by the searches of the intranet mediator.

“Common schema” means an organizational definition of structured data shared by multiple data sources.

15 “Data source” means a logically, independently operating data storage, search, retrieval, and manipulation system. The system may store data of any digital form.

SUMMARY OF THE INVENTION

The intranet mediator of the present invention obtains for the user a direct, or most likely, answer to a natural language question. The intranet mediator provides the user with a search of both structured data sources and a repository of unstructured data sources. The structured data sources will preferably be aggregated

into a physical data warehouse to provide for ease of searching a rapid attainment of search results. The data warehouse was previously constructed, e.g., by the corporate owner thereof, to integrate and abstract its collections of structured data sources with common schema, thereby providing meta-data for the data warehouse. The meta-data will give a global overview of the structured data sources thus making the warehouse easily searchable. The unstructured data sources may be data such as internal video, audio, or document files stored within the private databases of the owner, or they may be public sources, e.g., collections of documents available via the worldwide web or the like, or both. The unstructured data sources may be provided with a meta-data repository also.

The intranet mediator allows the user to input, and obtain an answer to, a natural language question without having to read text or surf the data sources in which the answer might be contained, or without being limited to one specific factual item return. The intranet mediator operates in part on the supposition that most answers to a businessperson's questions are contained within privately owned structured data sources of the business that were already integrated into a data warehouse. Selection of the most relevant data sources prior to searching is therefore possible. The direct answer selection to the user's question is accordingly weighted to the search response from these structured data sources. Searching of the unstructured data sources may be performed automatically, or upon determination that answers are not likely from the structured data sources, or if additional context surrounding either the query or the answer is justified or desired. A direct, or most

likely, answer is then given to the user in response to the input question. If desired, the intranet mediator may also display a list of related data or data sources where additional information relevant to the user's question may be found.

The intranet mediator desirably includes the logical functionality of a natural language question input module; a parser module for assembling a search query from the natural language question; a query expander module; an unstructured data source manager operably connected to at least one unstructured data source; a data source selection module operably connected to a meta-data repository; a dispatcher module interfacing between the data source selection module and both a structured data source manager and the unstructured data source manager, the structured data source manager being operably connected to a data warehouse; a results manager module operably connected to both of the data source managers; and an answer output module.

Discussion of the modules will be given herein with respect to specific functional tasks or task groupings that are in some cases arbitrarily assigned to the specific modules for explanatory purposes. It will be appreciated by the person having ordinary skill in the art that an intranet mediator according to the present invention may be arranged in a variety of ways, or that functional tasks may be grouped according to other nomenclature or architecture than is used herein without doing violence to the spirit of the present invention.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a block diagram representing the architecture of an intranet mediator according to one embodiment of the current invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

5 Referencing the block diagram of Fig. 1, the preferred embodiment of an intranet mediator 11 according to the present invention comprises means for the searching of both structured data sources, such as a multiplicity of databases integrated into a data warehouse 13, and unstructured data sources, collectively 15, to arrive at a direct answer to a natural language question input by the user.

10 A user interface 17, such as a known graphical user interface, accepts input of a user question in natural language format and may allow the user to manually select data sources if desired. As indicated by the two boxes in Fig. 1 labeled 17, the user interface 17 is preferably a part of an input/output user interface which is also tasked with displaying an answer to the user question.

15 A parser module 19 then accepts the natural language question input by the user, as at line 20, parsing and assembling the relevant concepts of the natural language question into a query, or queries, of weighted search tokens, hereinafter referred to simply as a query, and also eliminating the irrelevant, or non-indicative, words of the natural language question from use as search tokens. For example, in the
20 query: "What are the three best Sushi restaurants in Chicago?", words irrelevant to data retrieval such as "are" and "the" may be eliminated from use as tokens by the parser module. Various techniques such as grammar matching, lattices, partial

lattices, etc. for accomplishing the tasks of the parser module 19 are available and considered within the skill of the person having ordinary skill in the art.

A query expander module 21 is operably connected to the parser module 19, as at line 22, for receiving the parsed query tokens and obtaining additional terms synonymous or analogous to the tokens and expanding the query with additional tokens if desirable to expand the chance of a return of information relevant to the query. Various techniques such as thesauri, dictionaries, term expansion, stemming, phrase generation, and the like are available to accomplish this query, or token, expansion and are considered within the skill of the person having ordinary skill in the art.

The query expander module 21 then passes the tokens, preferably including the expanded tokens, to an unstructured data source manager 23, as at line 24, for a limited use specialty search to acquire additional synonyms, or terms analogous to, the query tokens and expanded query tokens, if any. The unstructured data source manager 23 initiates a search, as at lines 26, of one or more repositories of unstructured data sources 15, hereinafter referred to in the singular for ease of explanation, and obtains the search results therefrom, as at lines 28, with any acceptable method of information retrieval. Additional tokens, if any, may be filtered from the results of the unstructured data source search returned to the query expander module 21 from unstructured data source manager 23 via line 46 and added to the expanded query for return to the parser module 19 through the query expander module 21, as at line 30. The unstructured data repository 15, or an operable link to such a

repository, may also be considered as a part of the intranet mediator 11 according to certain embodiments of the present invention. Unstructured data sources may include private or public data repositories, e.g., a private video archive and the World Wide Web, respectively.

5 The parser module 19 will then pass the expanded query to the data source selection module 25, as at line 32. The data source selection module 25 determines the most likely data source to contain an answer for each token of the expanded query; e.g., adaptive, heuristic, hard-coded, user-directed, standard, handwritten, or data-mined inquiry techniques through its connection, as at line 34,
10 to a meta-data repository 27 abstracting the contents of a data warehouse 13 of structured data sources. Meta-data for the unstructured data repository 15 may further be contained in the meta-data repository 27 in certain embodiments of the present invention. The meta-data repository, or an operable link to such a repository, may also be considered as a part of the intranet mediator according to certain embodiments of
15 the present invention.

Upon selection of the appropriately likely data sources to return an answer for each token of the expanded query, the data source selection module 25 will pass the data source selections and accompanying search tokens onto the dispatcher module 31, as at line 36. The dispatcher module 31 will then route each token to be searched to the appropriate data source manager 33, 23. Routing will commonly occur via an intranet although other networks including the internet or simply direct routing are also within the scope of this invention, as at line 38 to the structured data

source manager 33 or to the unstructured data source manager 23 via line 40. The structured data source manager 33 and the unstructured data source manager 23 will then interface most efficiently with the data warehouse or each database, via lines 42 and 26, respectively, according to the individual requirements of the data sources.

5 The data warehouse 13 is a physical data warehouse of integrated and structured private data sources and may be constructed according to known techniques such as extraction/transform/load (ETL) techniques. One of ordinary skill in the art will recognize that other data warehouse development techniques are likewise within the scope of this invention. The data warehouse 13, or an operable link to such a warehouse, may also be considered a part of the intranet mediator 11 according to 10 some embodiments of the present invention.

Search results are returned to the structured and unstructured data source managers, 33, 23 respectively, via lines 44, 28, respectively, for additional processing if any, and then forwarded via lines 46, 48 respectively to the results manager module

15 51.

The results manager module 51 desirably accepts and consolidates the results of the structured and unstructured data source searches for each search token and integrates the results of the searches. Duplicate results are also eliminated as appropriate. The results manager module 51 then weights or ranks the results and at 20 least one most likely answer is selected. In the presently preferred embodiment, if an answer is returned from the structured data source, for example, the data warehouse, it is published as the most likely answer. It will be understood that selection of the

direct answer from the unstructured data source is within the scope of this invention. In the described embodiment, the associated data links, extracted text, or the like, from the unstructured data source search may be further selected and ranked. An answer module portion 53 of the user interface 17 may then receive answer data via 5 line 50 and may then perform any formatting necessary for the display via the user interface 17, of the most likely answer and, if desired, the associated data or data links available as additional context for the direct answer.

The user interface 17 of the described embodiment will wait for some specified time interval or specified number of results to accumulate and then display 10 the currently ranked results. The user interface 17 will further continue receiving results and ranking them in conjunction with the total accumulated results to capture answers which may have been delayed for various reasons. The user is then preferably presented with an option to activate a "get next" display command, and the now possibly revised rankings or additionally aggregated results are displayed. 15

Having thus described an intranet mediator for searching both structured and unstructured data sources for arriving at a most likely answer to a natural language question input by the user; it will be appreciated that many variations thereon will occur to the artisan upon an understanding of the present invention, which is therefore to be limited only by the appended claims.